Using Soft Constraints in Joint Inference for Clinical Concept Recognition

Prateek Jindal and Dan Roth
Department of Computer Science, UIUC
201 N. Goodwin Ave, Urbana, IL 61801, USA
{jindal2, danr}@illinois.edu

Abstract
This paper introduces IQPs (Integer Quadratic Programs) as a way to model joint inference for the task of concept recognition in clinical domain. IQPs make it possible to easily incorporate soft constraints in the optimization framework and still support exact global inference. We show that soft constraints give statistically significant performance improvements when compared to hard constraints.

1 Introduction
In this paper, we study the problem of concept recognition in the clinical domain. State-of-the-art approaches (de Bruijn et al., 2011; Patrick et al., 2011; Torii et al., 2011; Minard et al., 2011; Jiang et al., 2011; Xu et al., 2012; Roberts and Harabagiu, 2011; Jindal and Roth, 2013) for concept recognition in clinical domain (Uzuner et al., 2011) use sequence-prediction models like CRF (Lafferty et al., 2001), MEMM (McCallum et al., 2000) etc. These approaches are limited by the fact that they can model only local dependencies (most often, first-order models like linear chain CRFs are used to allow tractable inference).

Clinical narratives, unlike newswire data, provide a domain with significant knowledge that can be exploited systematically to improve the accuracy of the prediction task. Knowledge in this domain can be thought of as belonging to two categories: (1) Background Knowledge captured in medical ontologies like UMLS (U11, 2013), MeSH and SNOMED CT and (2) Discourse Knowledge driven by the fact that the narratives adhere to a specific writing style. While the former can be used by generating more expressive knowledge-rich features, the latter is more interesting from our current perspective, since it provides global constraints on what output structures are likely and what are not. We exploit this structural knowledge in our global inference formulation.

Integer Linear Programming (ILP) based approaches have been used for global inference in many works (Roth and Yih, 2004; Punyakanok et al., 2004; Punyakanok et al., 2008; Marciniak and Strube, 2005; Bramsen et al., 2006; Barzilay and Lapata, 2006; Riedel and Clarke, 2006; Clarke and Lapata, 2007; Clarke and Lapata, 2008; Denis et al., 2007; Chang et al., 2011). However, in most of these works, researchers have focussed only on hard constraints while formulating the inference problem.

Formulating all the constraints as hard constraints is not always desirable because the constraints are not perfect in many cases. In this paper, we propose Integer Quadratic Programs (IQPs) as a way of formulating the inference problem. IQPs is a richer family of models than ILPs and it enables us to easily incorporate soft constraints into the inference procedure. Our experimental results show that soft constraints indeed give much better performance than hard constraints.

2 Identifying Medical Concepts

Task Description Our input consists of clinical reports in free-text (unstructured) format. The task is: (1) to identify the boundaries of medical concepts and (2) to assign types to such concepts. Each concept can have 3 possible types: (1) Test, (2) Treatment, and (3) Problem. We would refer to these three types by TEST, TRE and PROB in the following discussion.

Our Approach In the first step, we identify the concept boundaries using a CRF (with BIO encod-
[Chest x-ray] gave positive evidence for [atelectasis] and [sarcoidosis].

(a) Example 1

No [hemoptysis], [hematemesis], [urgency], [abdominal pain], [black or tarry stools], [dysuria].

(b) Example 2

Figure 1: This figure motivates the global inference procedure we used. For discussion, please refer to §2.

Inference Procedure: The final assignment of types to concepts is determined by an inference procedure. The basic principle behind our inference procedure is: “Types of concepts which appear close to one another are often closely related. For some concepts, type can be determined with more confidence. And relations between concepts’ types guide the inference procedure to determine the types of other concepts.” We will now explain it in more detail with the help of examples. Figure 1 shows two sentences in which the concepts are shown in brackets and correct (gold) types of concepts are shown above them.

First, consider first and second concepts in Figure 1a. These concepts follow the pattern: [Concept1] gave positive evidence for [Concept2]. In clinical narratives, such a pattern strongly suggests that Concept1 is of type TEST and Concept2 is of type PROB. Table 1 shows additional such patterns. Next, consider different concepts in Figure 1b. All these concepts are separated by commas and hence, form a list. It is highly likely that such concepts should have the same type.

3 Modeling Global Inference

Inference is done at the level of sentences. Suppose there are \( m \) concepts in a sentence. Each of the \( m \) concepts has to be assigned one of the following types: TEST, TRE, PROB or NULL. To represent this as an inference problem, we define the indicator variables \( x_{i,j} \) where \( i \) takes values from 1 to \( m \) (corresponding to concepts) and \( j \) takes values from 1 to 4 (corresponding to 4 possible types). \( p_{i,j} \) refers to the probability that the \( i \)th concept has type \( j \).

We can now write the following optimization problem to find the optimal concept types:

\[
\max_x \sum_{i=1}^{m} \sum_{j=1}^{4} x_{i,j} \cdot p_{i,j}
\]

subject to

\[
\sum_{j=1}^{4} x_{i,j} = 1 \quad \forall i
\]

\[
x_{i,j} \in \{0, 1\} \quad \forall i, j
\]

The objective function in Equation (1) expresses the fact that we want to maximize the expected number of correct predictions in each sentence. Equation (2) enforces the constraint that each concept has
a unique type. We would refer to these as Type-1
constraints.

3.1 Constraints Used
In this subsection, we will describe two addi-
tional types of constraints (Type-2 and Type-3) that were added to the optimization procedure de-
scribed above. Whereas Type-1 constraints de-
scribed above were formulated as hard constraints,
Type-2 and Type-3 constraints are formulated as
soft constraints.

3.1.1 Type-2 Constraints
Certain constructs like comma, conjunction, etc.
suggest that the 2 concepts appearing in them should
have the same type. Figure 1b shows an example
of such a constraint. Suppose that there are \( n_2 \) such constraints. Also, assume that the \( i^{th} \) constraint says that the concepts \( R_i \) and \( S_i \) should have the same

\[
\min \sum_{m=1}^{4} (x_{R_i,m} - x_{S_i,m})^2
\]

subject to

\[
x_{i,j} = 1 \quad \forall i
\]

\[
z_k = 1 \iff x_{A_{i,k},B_{1,k}} \land x_{A_{i,k},B_{2,k}} \forall k \in \{1...n_3\}
\]

Figure 2: Final Optimization Problem (an IQP)

After incorporating all the constraints mentioned
above, the final optimization problem (an IQP) is
shown in Figure 2. We used Gurobi toolkit (Url6,
2013) to solve such IQPs. In our case, it solves
76 IQPs per second on a quad-core server with Intel
Xeon X5650 @ 2.67 GHz processors and 50 GB
RAM.

4 Experiments and Results

4.1 Datasets and Evaluation Metrics
For our experiments, we used the datasets pro-
vided by i2b2/VA team as part of 2010 i2b2/VA
shared task (Uzuner et al., 2011). The datasets
used for this shared task contained de-identified clinical
reports from three medical institutions: Partners
Healthcare (PH), Beth-Israel Deaconess Medical
Center (BIDMC) and the University of Pitts-
burgh Medical Center (UPMC). UPMC data was di-
vided into 2 sections, namely discharge
(UPMC) and progress notes
(UPMCP). A total of 349 training
reports and 477 test reports were made available
to the participants. However, data which came from
UPMC (more than 50% data) was not made avail-
able for public use. As a result, we had only 170
clinical reports for training and 256 clinical reports
for testing. Table 3 shows the number of clinical re-
ports made available by different institutions. The

\[
\min \sum_{m=1}^{4} (x_{R_i,m} - x_{S_i,m})^2
\]

subject to

\[
x_{i,j} = 1 \quad \forall i
\]

\[
z_k = 1 \iff x_{A_{i,k},B_{1,k}} \land x_{A_{i,k},B_{2,k}} \forall k \in \{1...n_3\}
\]
Table 2: Our final system, BKC, consistently performed the best among all 4 systems (B, BK, BC and BKC).

<table>
<thead>
<tr>
<th>PH</th>
<th>BIDMC</th>
<th>UPMCD</th>
<th>UPMCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>97</td>
<td>73</td>
<td>98</td>
</tr>
<tr>
<td>Test</td>
<td>133</td>
<td>123</td>
<td>402</td>
</tr>
</tbody>
</table>

Table 3: Dataset Characteristics

<table>
<thead>
<tr>
<th></th>
<th>PH</th>
<th>BIDMC</th>
<th>UPMCD</th>
<th>UPMCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>97</td>
<td>73</td>
<td>98</td>
<td>84</td>
</tr>
<tr>
<td>Test</td>
<td>133</td>
<td>123</td>
<td>402</td>
<td>449</td>
</tr>
</tbody>
</table>

4.2 Results

In this section, we would refer to following 4 systems: (1) Baseline (B), (2) Baseline + Knowledge (BK), (3) Baseline + Constraints (BC) and (4) Baseline + Knowledge + Constraints (BKC). Please note that the difference between B and BK is that B does not use the features derived from domain-specific knowledge sources (namely MetaMap, UMLS, MeSH and SNOMED CT) for training the classifiers. Both B and BK do not use the inference procedure. BKC uses all the features and also the inference procedure. In addition to these 4 systems, we would refer to another system, namely, BKC-HARD. This is similar to BKC system. However, it sets $\rho_2 = \rho_3 = 1$ which effectively turns Type-2 and Type-3 constraints into hard constraints by imposing very high penalty.

4.2.1 Importance of Soft Constraints

Figures 3a and 3b show the effect of varying the penalties ($\rho_2$ and $\rho_3$) for Type-2 and Type-3 constraints respectively. These figures show the F1-score of BKC on the development set. Penalty of 0 means that the constraint is not active. As we increase the penalty, the constraint becomes stronger. As the penalty becomes 1, the constraint becomes hard in the sense that final assignments must respect the constraint. We observe from Figures 3a and 3b that for Type-2 and Type-3 constraints, global maxima is attained at $\rho_2 = 0.6$ and $\rho_3 = 0.3$ respectively.

**Hard vs Soft Constraints** Table 4 compares the performance of BKC-HARD with that of BKC. First 3 rows in this table show the performance of both systems for the individual categories (TEST, TRE and PROB). The fourth row shows the overall score of both systems. BKC outperformed BKC-HARD on all the categories by statistically significant differences at $p = 0.05$ according to Bootstrap Resampling Test (Koehn, 2004). For the OVERALL category, BKC improved over BKC-HARD by 1.0 F1 points.

<table>
<thead>
<tr>
<th></th>
<th>BKC-HARD</th>
<th>BKC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST</td>
<td>84.7</td>
<td>85.8</td>
</tr>
<tr>
<td>TRE</td>
<td>84.7</td>
<td>85.7</td>
</tr>
<tr>
<td>PROB</td>
<td>85.6</td>
<td>86.7</td>
</tr>
<tr>
<td>OVERALL</td>
<td>85.1</td>
<td>86.1</td>
</tr>
</tbody>
</table>
4.2.2 Comparing with state-of-the-art baseline

In the 2010 i2b2/VA shared task, majority of top systems were CRF-based models, motivating the use of CRF as our baseline. Table 2 compares the performance of 4 systems: B, BK, BC and BKC. As pointed out before, our BK system uses CRF for boundary detection, employs all the knowledge-based features and is very similar to the top-performing systems in i2b2 challenge. We see from Table 2 that BKC consistently performed the best for individual as well as overall categories. This result is statistically significant at $p = 0.05$ according to Bootstrap Resampling Test (Koehn, 2004). It should also be noted that BC performed significantly better than B for all the categories. Thus, the constraints are helpful even in the absence of knowledge-based features. Since we report results on publicly available datasets, future works would be able to compare their results with ours.

4.2.3 Effect of training data size

In Figure 4, we report the overall F1-score on a part of the development set as we vary the size of the training data from 40 documents to 130 documents. We notice that the performance increases steadily as more and more training data is provided. This suggests that if we could train on full training data as was made available in the challenge, the final scores could be much higher. We also notice from the figure that BKC consistently outperforms the state-of-the-art BK system as we vary the size of the training data, indicating the robustness of the joint inference procedure.

1 Please note that the results reported in Table 2 can not be directly compared with those reported in the challenge because we only had a fraction of the original training and testing data.

5 Discussion and Related Work

In this paper, we chose to train a rather simple sequential model (using CRF), and focused on incorporating global constraints only at inference time. While it is possible to jointly train the model with the global constraints (as illustrated by Chang et al. (2007), Mann and McCallum (2007), Mann and McCallum (2008), Ganchev et al. (2010) etc.), this process will be a lot less efficient, and prior work (Roth and Yih, 2005) has shown that it may not be beneficial.

Roth and Yih (2004, 2007) suggested the use of integer programs to model joint inference in a fully supervised setting. Our paper follows their conceptual approach. However, they used only hard constraints in their inference formulation. Chang et al. (2012) extended the ILP formulation and used soft constraints within the Constrained Conditional Model formulation (Chang, 2011). However, their implementation performed only approximate inference. In this paper, we extended the integer linear programming to a quadratic formulation, arguing that it simplifies the modeling step, and showed that it is possible to do exact inference efficiently.

Conclusion

This paper presented a global inference strategy (using IQP) for concept recognition which allows us to model structural knowledge of the clinical domain as soft constraints in the optimization framework. Our results showed that soft constraints are more effective than hard constraints.

Acknowledgments

This research was supported by Grant HHS 90TR0003/01 and by IARPA FUSE program via DoI/NBC contract #D11PC2015. Its contents are solely the responsibility of the authors and do not necessarily represent the official views, either expressed or implied, of the HHS, IARPA, DoI/NBC or the US government. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

In another experiment, we replaced the CRF with an MEMM. Surprisingly, MEMM performed as well as CRF.

It should be noted that it is possible to reduce IQPs to ILPs using variable substitution. However, the resulting ILPs can be exponentially larger than original IQPs.
References


